Grammar-driven versus Data-driven: Which Parsing System is More Affected by Domain Shifts?

Barbara Plank, Gertjan van Noord
University of Groningen, The Netherlands

July 16, 2010
NLPling 2010 Workshop, Uppsala, Sweden

Motivation

- Past decade: development of various systems for parsing natural language, based on different parsing approaches

Motivation

- Past decade: development of various systems for parsing natural language, based on different parsing approaches
- What they have in common: problem of lack of portability to new domains, i.e. $\longrightarrow$ drop in performance when tested on data from another domain

- E.g., for PCFG Parsing, English:

	Test			
Train	WSJ	Brown	Genia	ETT
WSJ	89.7	84.1	76.2	82.2

Table: F-scores, Charniak parser (McClosky, Charniak & Johnson, 2010)

Motivation

Work on Domain Adaptation for Parsing

- Most research has been done for statistical systems (Gildea, 2001; Roark and Bacchiani, 2003; McClosky et al., 2006; Dredze et al., 2007)
- Little work on adaptation of grammar-based (hand-crafted) parsing systems (Hara, 2005; Plank and van Noord, 2008)

Motivation

Work on Domain Adaptation for Parsing

- Most research has been done for statistical systems (Gildea, 2001; Roark and Bacchiani, 2003; McClosky et al., 2006; Dredze et al., 2007)
- Little work on adaptation of grammar-based (hand-crafted) parsing systems (Hara, 2005; Plank and van Noord, 2008)

Is the problem the same for different kind of parsing systems?

- **Hypothesis:** Grammar-driven systems are less affected by domain changes.

Motivation

Work on Domain Adaptation for Parsing

- Most research has been done for statistical systems (Gildea, 2001; Roark and Bacchiani, 2003; McClosky et al., 2006; Dredze et al., 2007)
- Little work on adaptation of grammar-based (hand-crafted) parsing systems (Hara, 2005; Plank and van Noord, 2008)

Is the problem the same for different kind of parsing systems?

- **Hypothesis:** Grammar-driven systems are less affected by domain changes.

Empirical Investigation on Dutch

- Evaluate different dependency parsing systems across Domains
- Propose a simple measure to quantify domain sensitivity

Parsers

Grammar-driven

- **Alpino**
  - Parser for Dutch (hand-crafted HPSG grammar)
  - Developed over last 10 years (from a domain-specific HPSG)
  - 800 rules, large hand-crafted lexicon, unknown word heuristics, left-corner parser
  - Separate statistical disambiguation component (MaxEnt)

Data-driven

- **MST parser (MST)**
  - Graph-based dependency parser
- **Malt parser (Malt)**
  - Transition-based dependency parser

Datasets

- **Train data - Source: Newspaper text**
  - Alpino Treebank (cdb): 7,136 sentences from Eindhoven corpus
  - collection of text fragments from 6 Dutch newspapers
- **Test data - Target:**
  1. Wikipedia
     - 95 Dutch Wikipedia articles which were annotated in the course of the LASSY project
     - Mostly about Belgium issues, i.e. locations, politics, etc.
     - 10 subdomains
  2. DPC (Dutch Parallel Corpus)
     - 186 articles
     - 13 subdomains (a.o.: medical, oceanography, etc.)

Parser Performance Across Domains

- Which parsing system is more affected by domain shifts?
- Or: ... more robust to different input texts?

$\rightarrow$ Robustness in terms of performance variation

Parser Performance Across Domains

- Which parsing system is more affected by domain shifts?
- Or: ... more robust to different input texts?

$\rightarrow$ Robustness in terms of performance variation

Towards a Measure of Domain Sensitivity

- Intuitive measure: mean (μ) and standard deviation (sd) of performance on target domains LAS_p^i's
- Drawbacks:
  - sd highly influenced by outliers
  - does not take source domain (baseline) into consideration

Parser Performance Across Domains

Proposal: Average Domain Variation (adv)

$$adv = \sum_{i=1}^{N} w^i * \Delta_p^i$$

- Relative to source domain: $\Delta_p^i = LAS_p^i - LAS_p^{baseline}$
- Weighted by domain size: $w^i = \frac{size(w^i)}{\sum_{i=1}^{N} size(w^i)}$, with $\sum_{i=1}^{N} w^i = 1$

- Thus, *adv* can take on positive or negative values
  $\rightarrow$ to indicate average gain/loss w.r.t. baseline
- Or: *unweighted* variant (problem of threshold; in paper)

Experimental Setup

- Evaluation: Labeled Attachment Score (LAS)
- Baseline: 5-fold cross-validation on Alpino Treebank (cdb)

Experimental Setup

- Evaluation: Labeled Attachment Score (LAS)
- Baseline: 5-fold cross-validation on Alpino Treebank (cdb)

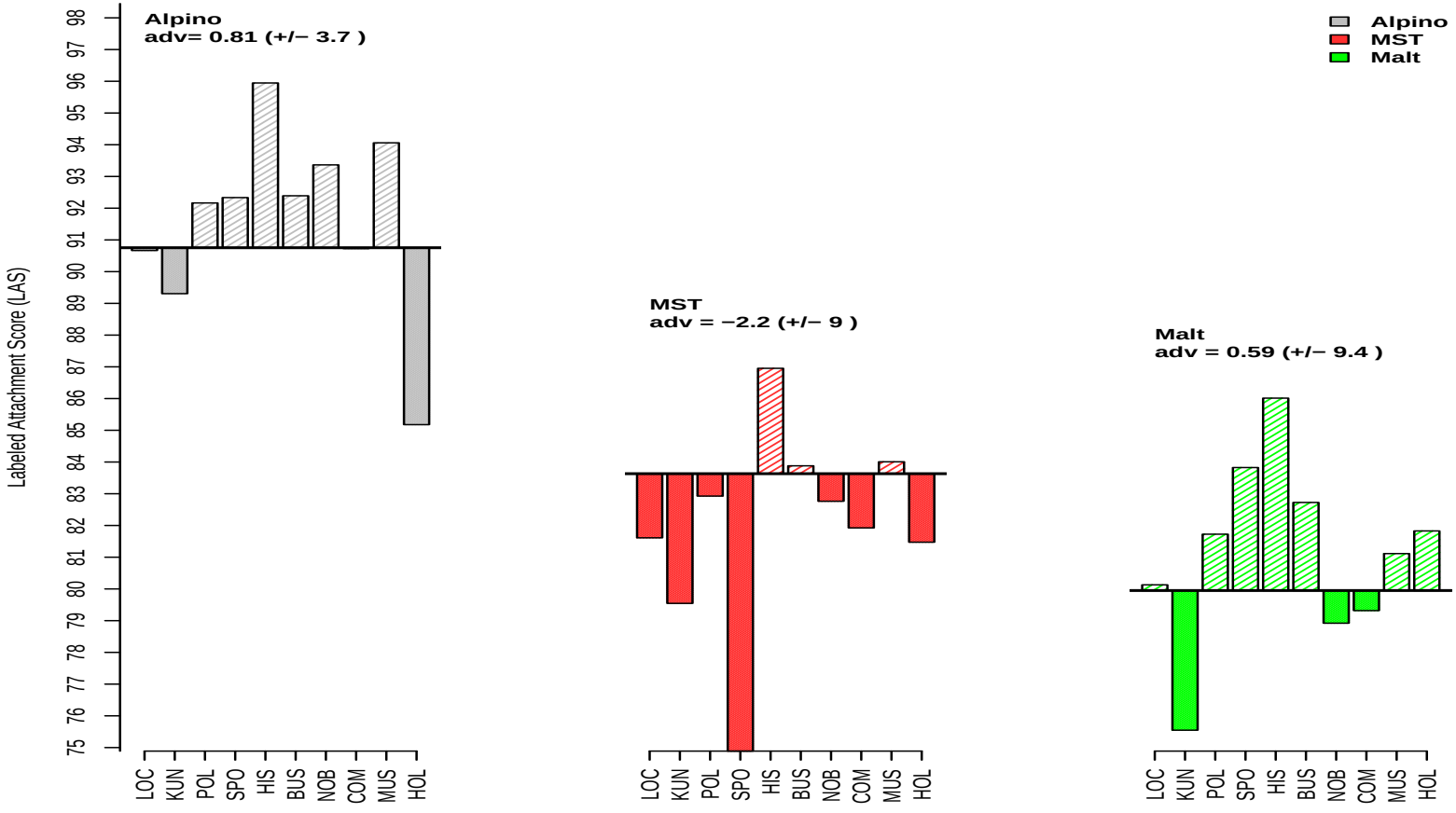
Training data-driven parsers - Sanity Checks

- Convert Alpino format to CoNLL (using Marsi's CoNLL conversion software, with PoS tags replaced by Alpino tags)
- Evaluated on CoNLL 2006 testset:

Model	**LAS**	**UAS**
MST (cdb retagged with Alpino)	82.14	85.51
Malt (cdb retagged with Alpino)	80.64	82.66
MST (state-of-the-art)	79.19	83.6
Malt (state-of-the-art)	78.59	n/a

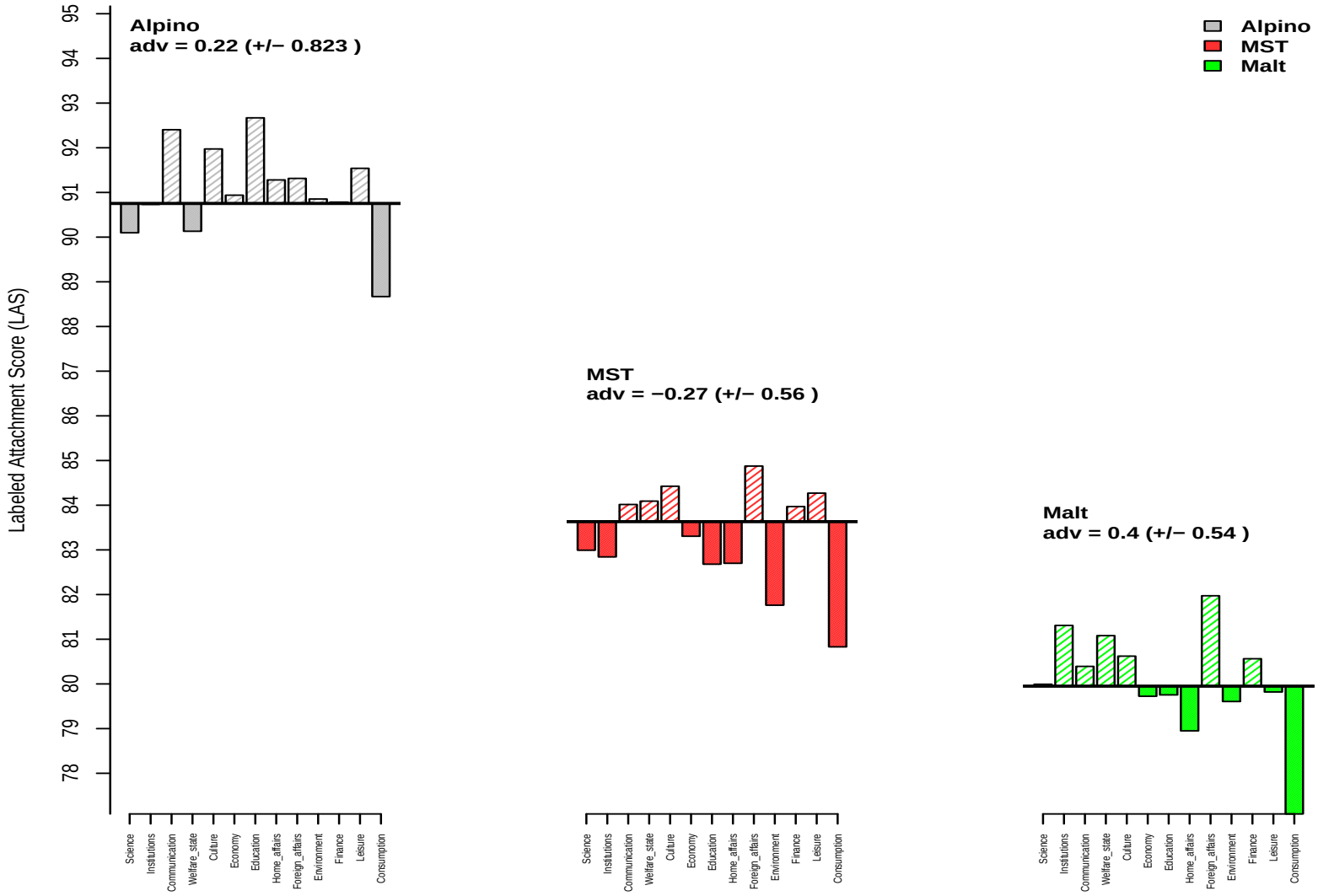
- Retagging helped (78.73 $\rightarrow$ 82.14)
- Despite standard settings and limited data, we are in line (and actually above) state-of-the-art

Evaluation (1) - Wikipedia



- Alpino does not suffer much (adv=0.81; often above baseline)
- MST suffers the most (adv = -2.2)
- Malt significantly worse (absolute), but less affected (adv=0.59) – graph similar if measured in UAS

Evaluation (2) - DPC



Evaluation - Discussion

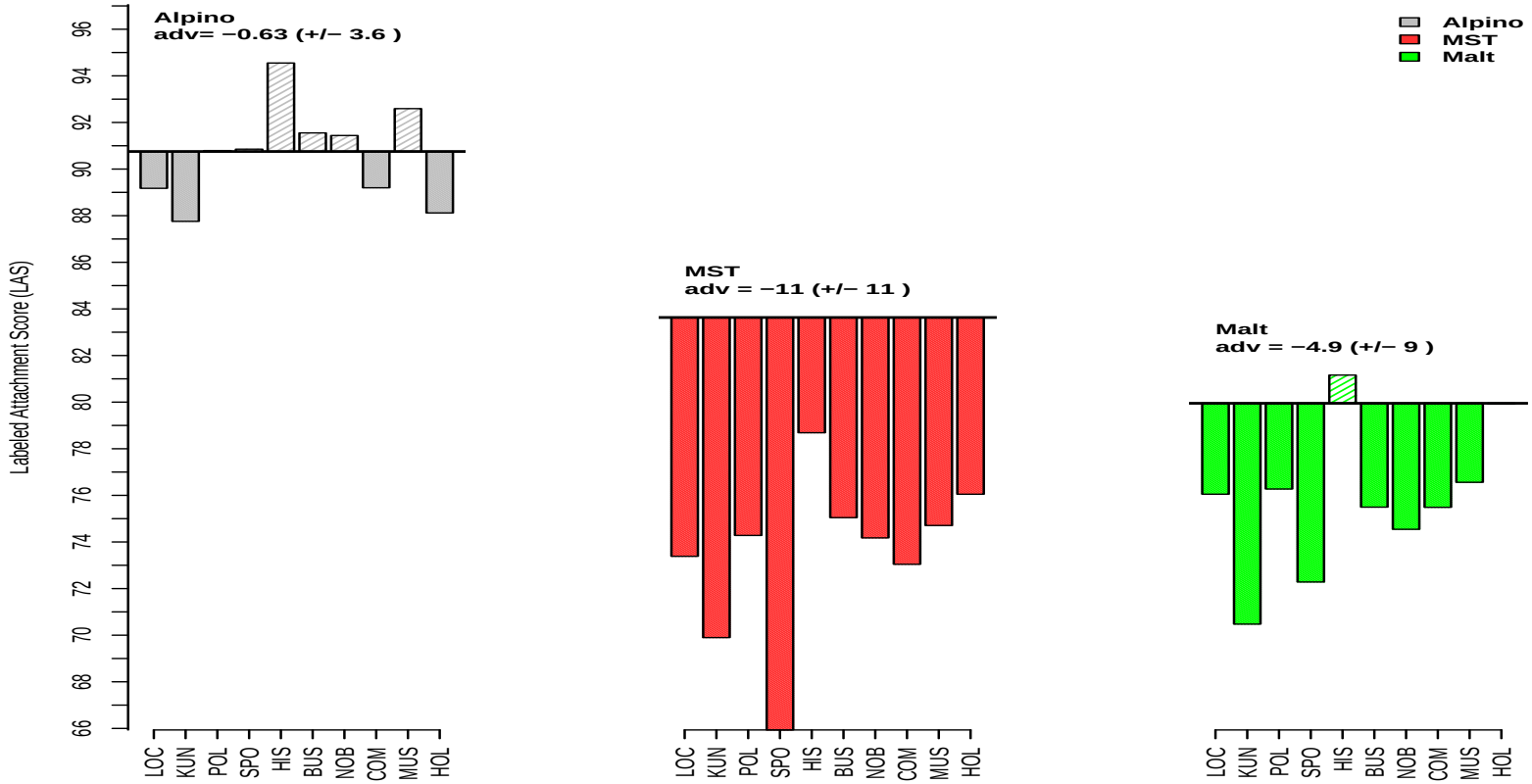
Are the differences significant?

- *Approximate Randomization Test* over 23 performance differences Δ_p^i (23 domains)
- Result:
  - $\Delta_{Alpino} \leftrightarrow \Delta_{MST}$: yes
  - $\Delta_{MST} \leftrightarrow \Delta_{Malt}$: yes
  - $\Delta_{Alpino} \leftrightarrow \Delta_{Malt}$: no

Excursion

- What happens when we exclude lexical information?

Evaluation (3) - Wikipedia (unlexicalized)



- Performance drops for all parsers in all domains
- As expected, for data-driven parsers to a higher degree

Conclusions and Future Work

Conclusions

- Examined domain sensitivity of different kind of parsing system for Dutch (data-driven versus grammar-driven)
- Proposed a simple measure to quantify domain sensitivity
- Results show that grammar-driven system Alpino is rather robust across domains; significantly more robust than MST

Future Work

- Perform error analysis (why for some domains parsers outperform baseline; what are typical in/out-domain errors)
- Examine why there is this difference between MST and Malt
- Investigate what part(s) of Alpino are responsible for differences with data-driven parsers

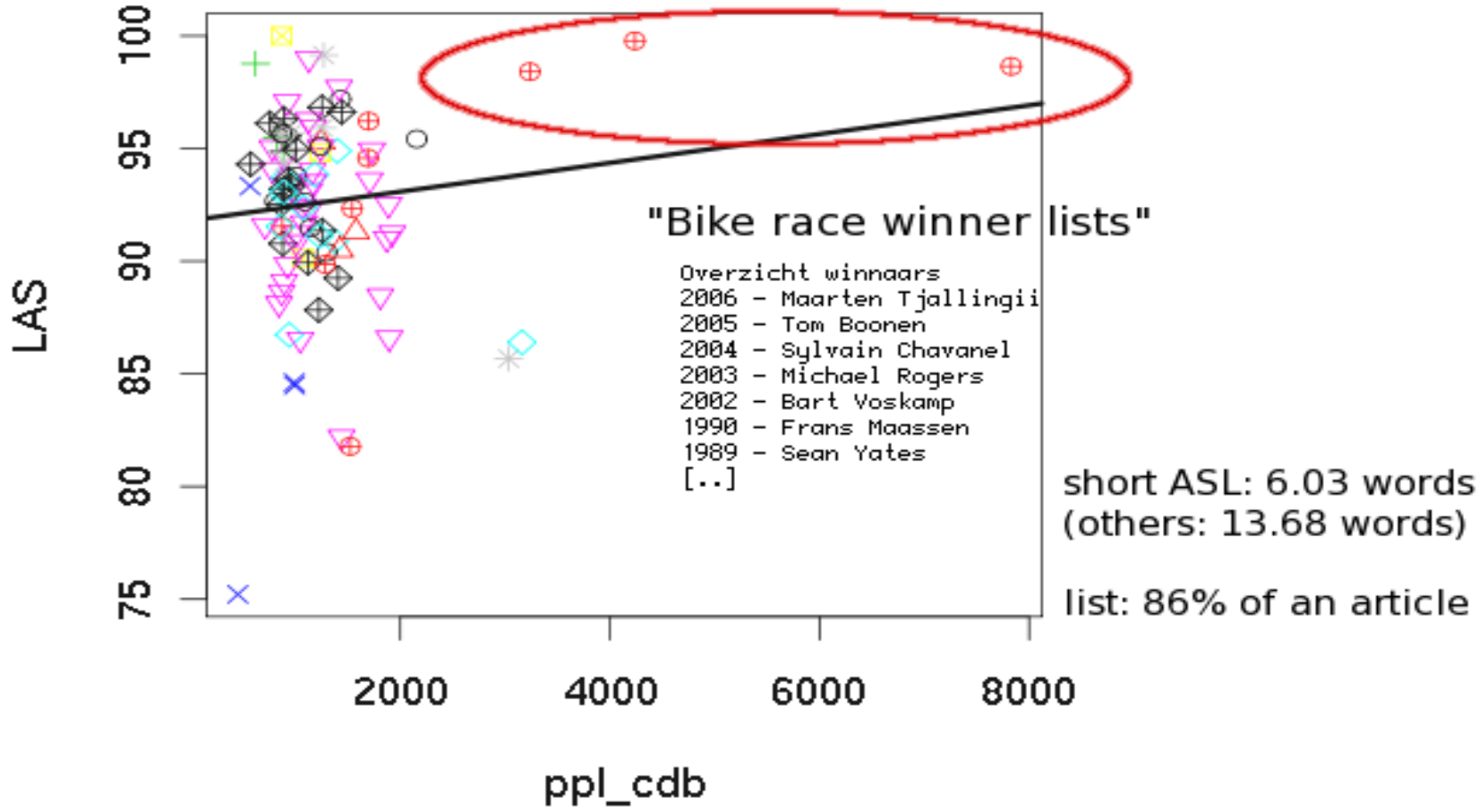
Questions? Comments? Suggestions?

Thank you!

Wikipedia sport articles

Parser performance against perplexity, per Wiki article

Alpino:



Alpino

sents	parses	oracle	arbitrary	model
536	45011	95.74	76.56	89.39

Wikipedia dataset

Wikipedia	Example articles	#a	#w	ASL
LOC (location)	Belgium, Antwerp (city)	31	25259	11.5
KUN (arts)	Tervuren school	11	17073	17.1
POL (politics)	Belgium elections 2003	16	15107	15.4
SPO (sports)	Kim Clijsters	9	9713	11.1
HIS (history)	History of Belgium	3	8396	17.9
BUS (business)	Belgium Labour Federation	9	4440	11.0
NOB (nobility)	Albert II	6	4179	15.1
COM (comics)	Suske and Wiske	3	4000	10.5
MUS (music)	Sandra Kim, Urbanus	3	1296	14.6
HOL (holidays)	Flemish Community Day	4	524	12.2
Total		**95**	**89987**	**13.4**

Table: Overview Wikipedia and DPC corpus (#a articles, #w words, ASL average sentence length)

Average number of sentences: 70.6
Average sentence length: 13.4
Cdb corpus Average sentence length: 19.7

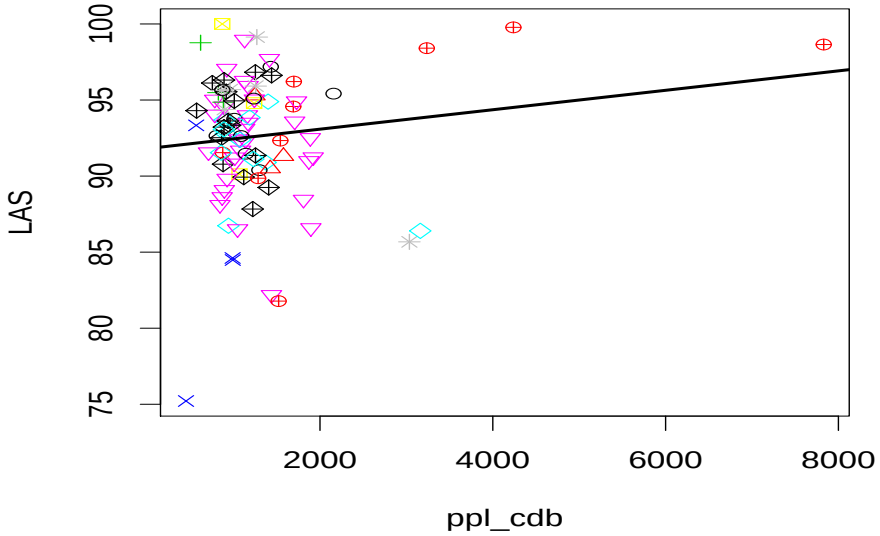
DPC

DPC	Description/Example	#a	#words	ASL
Science	medicine, oeanography	69	60787	19.2
Institutions	political speeches	21	28646	16.1
Communication	ICT/Internet	29	26640	17.5
Welfare state	pensions	22	20198	17.9
Culture	darwinism	11	16237	20.5
Economy	inflation	9	14722	18.5
Education	education in Flancers	2	11980	16.3
Home affairs	presentation (Brussel)	1	9340	17.3
Foreign affairs	European Union	7	9007	24.2
Environment	threats/nature	6	8534	20.4
Finance	banks (education banker)	6	6127	22.3
Leisure	various (drugscandal)	2	2843	20.3
Consumption	toys from China	1	1310	22.6
Total		**186**	**216371**	**18.5**

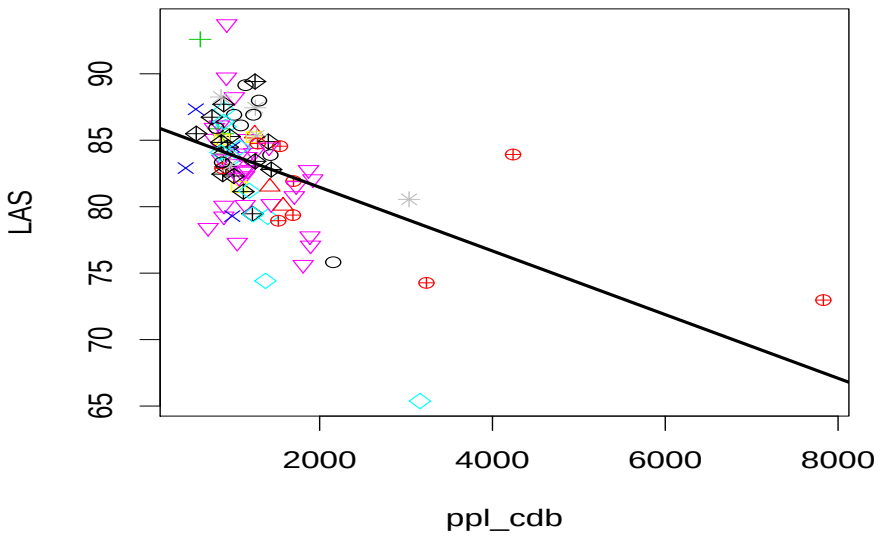
Table: Overview Wikipedia and DPC corpus (#a articles, #w words, ASL average sentence length)

Parser performance against perplexity, per Wiki article

Alpino (cor=0.1365)



MST (cor=-0.5124)

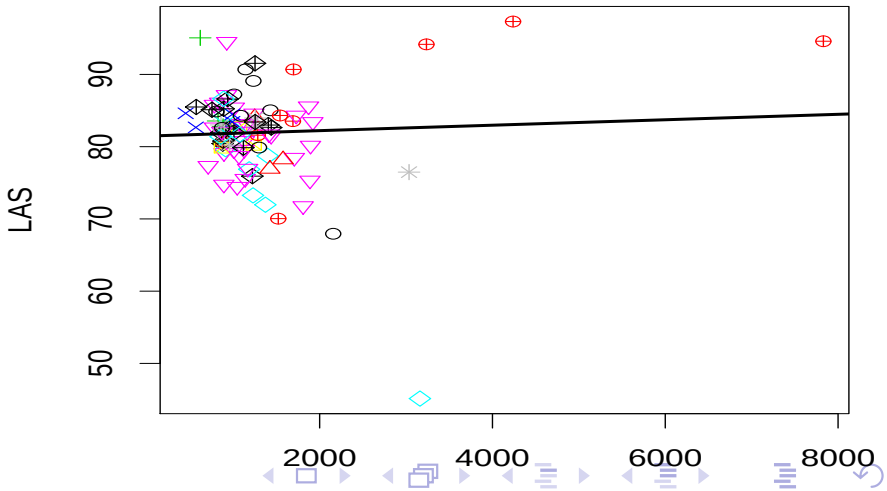


- 3 sports articles stand out (red crossed dots), cf. Plank, 2010

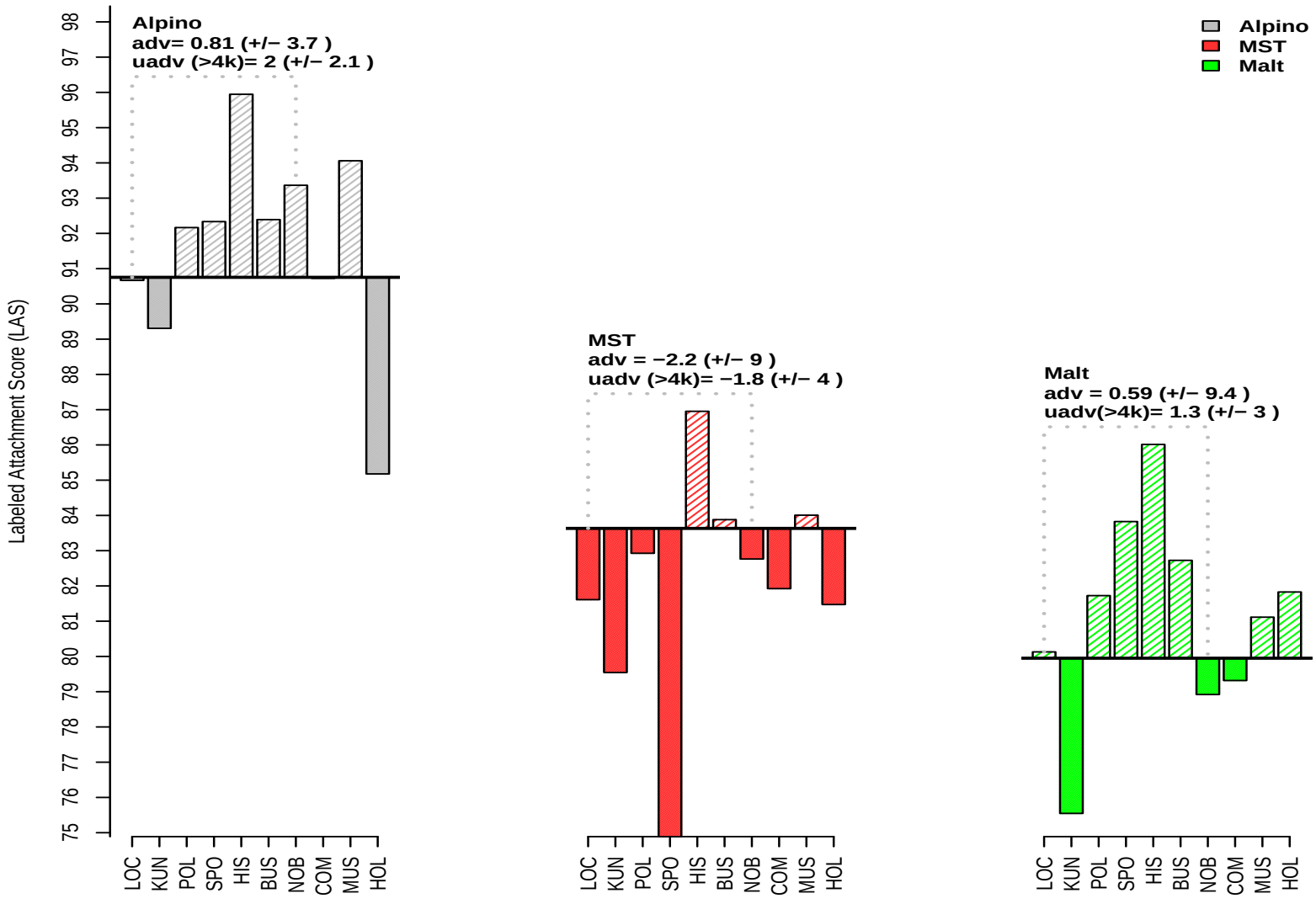
Legend:

BUS	LOC
COM	MUS
HIS	NOB
HOL	POL
KUN	SPO

Malt (cor=0.0520)



Extra slides: Wikipedia (unweighted)



Extra slides: DPC (unweighted)